

# Recognizing Local and Global Structural Motifs At the Atomic Scale

Piero Gasparotto, Robert Horst Meißner, and Michele Ceriotti\*

*Laboratory of Computational Science and Modeling, Institute of Materials, École  
Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland*

E-mail: michele.ceriotti@epfl.ch

## Abstract

Most of the current understanding of structure-property relations at the molecular and the supramolecular scales can be formulated in terms of the stability of and the interactions between a limited number of recurring structural motifs (e.g. H-bonds, coordination polyhedra, protein secondary-structure). Here we demonstrate an algorithm to automatically recognize such patterns, based on the identification of local maxima in the probability distributions observed in atomistic computer simulations, which is robust to the dimensionality and the sparsity of the reference atomistic data. We first discuss its main features, demonstrating some on artificial datasets, and then show how it can be applied to identify coordination environments in Lennard-Jones clusters, and to recognize secondary-structure patterns in the simulation of an oligopeptide. To assess the applicability of this algorithm for motifs that involve several interdependent degrees of freedom, we also employ it to identify groups of conformers of the cluster and the polypeptide, considered in their entirety. The motifs identified by analyzing atomistic simulations can be used to interpret and rationalize the stability and behavior of the system at hand, and also as a tool to accelerate sampling, in association with biased molecular dynamics schemes.

# 1 Introduction

The amount, availability and performance of computational facilities dedicated to atomic-scale modelling have increased dramatically in the last decades.<sup>1,2</sup> As a consequence, trajectories of several microseconds,<sup>3</sup> and systems comprising million of atoms<sup>4</sup> have become possible in atomistic simulations, and computationally-generated databases of materials properties have emerged as a tool to accelerate the design and discovery of new materials with improved properties.<sup>5-9</sup> The tremendous increase of the amount and complexity of data produced from simulations poses a challenge of its own, particularly when it comes to recognizing the molecular patterns that underlie the behavior of large, slowly-evolving systems. The traditional paradigm in which simulations were manually inspected, and analyzed in terms of structural descriptors designed by trial and error is becoming impractical: the amount and complexity of potential correlations concealed in such an amount of data are far from being fully understandable by a human.

A promising approach to facilitate the interpretation of large, complex databases and trajectories involves adapting established data-mining techniques to the analysis of atomic-scale data. For instance, one could apply clustering approaches to identify (meta)stable configurations, and partition the free-energy landscape in a set of fluctuation basins associated with each configuration. These basins can then be used as the basis for coarse-grained descriptions of the dynamics, such as Markov state models.<sup>10-13</sup> Many successful examples of machine-learning (ML)-based methods applied to the study of materials have also appeared in the literature in recent years.<sup>14-19</sup>

A particularly intriguing approach attempts to identify not only stable configurations of the system, but also the elementary structural motifs that, alone or combined in non-trivial ways, determine the stability and properties of complex supramolecular structures. This idea has seen application in the context of protein structure<sup>20-23</sup> materials<sup>24-26</sup> and molecular systems.<sup>27</sup> Recently the Probabilistic Analysis of Molecular Motifs (PAMM) algorithm has been developed with the precise goal of identifying molecular patterns based on an analysis

of the probability distribution of fragments observed in an atomistic simulation.<sup>28</sup> PAMM has this far been used to introduce a data-driven, agnostic definition of fundamental, yet simple, entities such as the hydrogen bond,<sup>28</sup> to recognize defects and correlations in liquid water,<sup>29</sup> and to single out protonated water species based on an analysis of their electronic structure.<sup>30</sup>

In this paper we address some of the limitations of the PAMM scheme, namely the stability with respect to dimensionality and sparse data sampling, and the reliability in case of periodic inputs or non-Gaussian features. In particular, we introduce an analysis of the stability of PAMM clusters that makes it possible to recognize the hierarchical structure of the free energy landscape. We show how the improved scheme can be used successfully not only to recognize local environments and molecular patterns, but also be extended to identify global structural features. In Section I we discuss the rationale and the practical implementation of our methodology. In Section II we apply our method to two realistic datasets. First, we identify coordination environments in a Lennard-Jones cluster and then recognize secondary-structure patterns in a  $\beta$ -hairpin oligopeptide. In Section III, as a demonstration of the stability of our approach when applied to high-dimensional descriptors, we use PAMM to perform a classification of entire structures, and show how clustering techniques can be combined with non-linear dimensionality reduction schemes to fully-characterize the configuration space of complex atomistic systems. Finally we draw our conclusions and discuss possible applications of probabilistic motif identifiers in the context of accelerated sampling schemes.

## 2 Probabilistic analysis of Molecular Motifs

The clustering algorithm we discuss in this paper follows the same philosophy of the probabilistic analysis of molecular motifs (PAMM) method,<sup>28</sup> including however a number of crucial additional ingredients to address the limitations of the original scheme. PAMM aims to provide an unambiguous and unbiased approach to recognize molecular motifs apparent

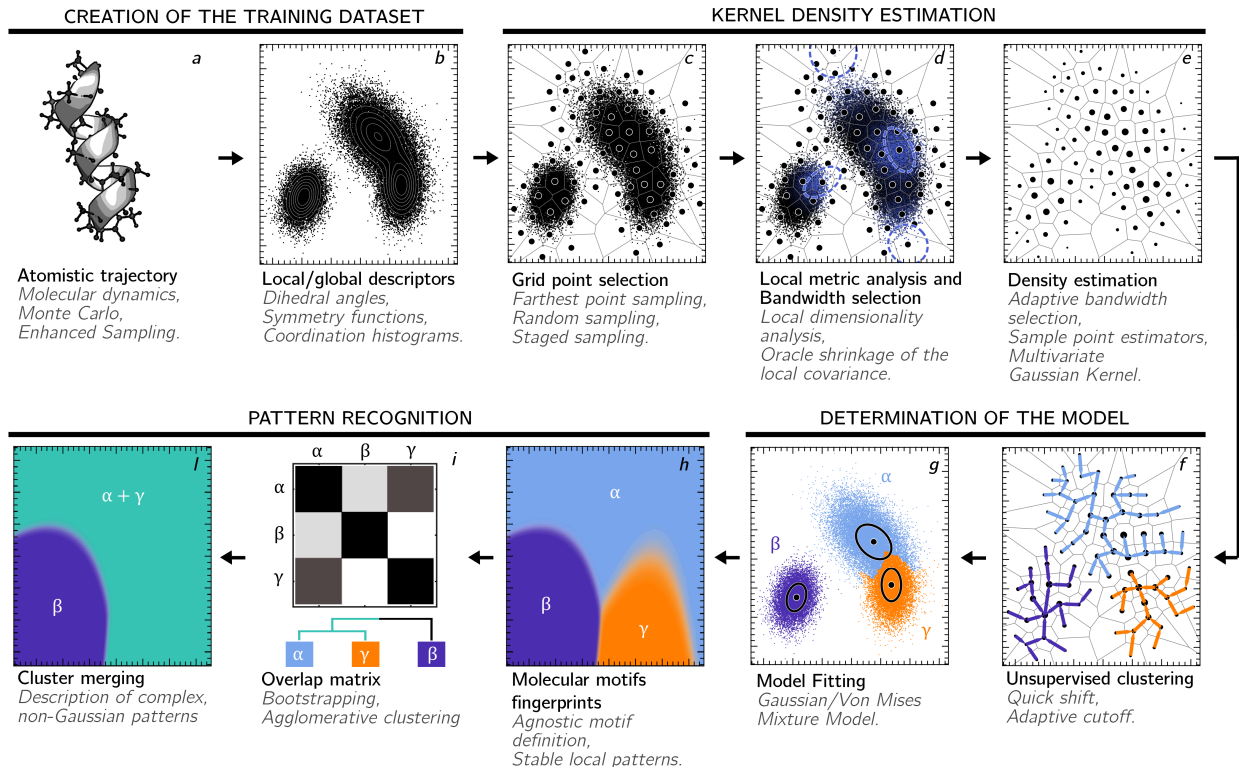


Figure 1: Schematic illustration of the PAMM workflow applied to an artificial two-dimensional dataset.

in data produced from atomistic simulations, by employing machine-learning techniques. Since the distribution of the structural descriptors reflects the (free) energetic stability of the molecular patterns, we use the probability distributions observed in the simulations to inform the identification of clusters. In line with the free-energetic interpretation, we choose to characterize distinct clusters as separate modes of the distribution, corresponding to the basins of attraction of local maxima.

Our approach can be split into several steps, that are summarized in Fig. 1. We will discuss separately the details of the algorithms we use for each step, which we found to be capable of providing robust clustering in a variety of realistic atomistic simulation problems. Different choices of course would be possible, as we will discuss when appropriate.

## 2.1 Input Representation

For most applications, it is impractical to directly use the Cartesian coordinates of the atoms to describe structures or local environments. One should rather represent them in terms of a (possibly large) number of “order parameters” or “fingerprints”, that provide an unbiased and sufficiently complete description of the geometry while fulfilling all of the important symmetries – such as being invariant to atom labelling or to rigid rotations and translations.<sup>31–33</sup> Seen through this lens, the atomistic simulation data is converted to a set  $\mathcal{X} = \{\mathbf{x}_i\}$  containing a large number  $N$  of  $D$ -dimensional vectors,  $\mathbf{x}_i \in \mathbb{R}^D$ , with each vector representing either a subset or the entirety of the atoms within a structure. It is worth stressing that the selection of a proper set of descriptors is far from being a trivial point, and can influence deeply the outcome of the subsequent analysis. To mitigate this problem, we optimized the different ingredients in PAMM to be robust with growing dimensionality, so that many order parameters can be used simultaneously to deal e.g. with heterogeneous systems. One could also use more abstract descriptors that are based on more or less systematic expansions of atomic environments – for instance Behler-Parrinello symmetry functions,<sup>34</sup> SOAP power spectra,<sup>32</sup> SPRINT coordinates.<sup>35</sup> The choice of input order parameters is also important, because it determines the metric relative to which probability distributions and free energies are computed.<sup>36,37</sup> A poor choice can generate spurious maxima in the probability distribution, or merge kinetically separate states into a single basin. These artifacts can be corrected, at least in part, by taking into account kinetic information in the definition of the order parameters.<sup>38,39</sup>

## 2.2 Grid Selection

In order to mitigate the high cost of density estimation for large datasets in high dimension, the first step in the PAMM workflow involves the selection of a sparse grid  $\mathcal{Y} \subseteq \mathcal{X}$  containing  $M \ll N$  points. A grid covering almost uniformly the parameter space spanned by  $\mathcal{X}$  can be obtained using a greedy farthest-point sampling procedure,<sup>40</sup> that can be modified to

use a finer grid in the high-probability regions.<sup>41</sup> The computational cost of this selection is  $\mathcal{O}(MN)$ , so it can be performed also on very large data sets. The determination of the grid by sub-sampling the full set  $\mathcal{X}$  also allows one to partition data into neighborhoods of the grid. For instance, one can construct the Voronoi polyhedra for  $\mathcal{Y}$ , and assign each datum  $x$  to the Voronoi set  $\mathcal{V}_i$  of the closest-by grid point  $y_i$ . Different strategies for subsampling are also possible.<sup>42,43</sup> It should be stressed, however, that the subsequent steps in the PAMM workflow are designed to minimize the impact of the grid size on the final outcome, and to guarantee that in the limit  $M, N \rightarrow \infty$  there is no dependence at all.

## 2.3 Kernel Density Estimation

Density-based clustering algorithms<sup>44,45</sup> depend crucially on the quality of the estimation of the underlying probability density. Kernel-density estimation (KDE) provides a smooth, robust approach to do so, that also leaves the flexibility to adapt to strongly anisotropic probability distributions and/or non-Euclidean geometries.

The KDE on a grid point  $\mathbf{y}_i$  can be written as

$$P(\mathbf{y}_i) = \frac{1}{\sum_{j=1}^N w_j} \sum_{j=1}^N w_j K_{\mathbf{H}_j}(\mathbf{x}_j - \mathbf{y}_i), \quad (1)$$

where we use a very general expression in which each data point can be assigned a weight  $w_i$  (e.g. to compensate for biased sampling), and an adaptive bandwidth matrix  $\mathbf{H}_j$ . The use of a grid implies that (for a fixed grid) the cost of evaluating  $P$  scales only linearly with the number of data points.

We use an anisotropic multivariate Gaussian kernel,

$$K_{\mathbf{H}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^D |\mathbf{H}|}} \exp \left[ -\frac{1}{2} \mathbf{x}^T \mathbf{H}^{-1} \mathbf{x} \right], \quad (2)$$

that provides enough flexibility to adapt to strong variations of the geometric distribution

of data points. A common problem with kernel density estimation, which is particularly severe for high-dimensional and/or sparsely populated datasets, is the optimization of the bandwidth of the kernel. The shape of the kernel, encoded in the bandwidth matrix  $\mathbf{H}$  in eq. (2), determines a trade-off between the statistical noise in the estimated density and a systematic error due to the smoothing of the true underlying density. When the true density  $P^*(\mathbf{x})$  is known, the optimal bandwidth can be selected by minimizing the Mean Integrated Squared Error (MISE,  $\int d\mathbf{x} [P^*(\mathbf{x}) - P(\mathbf{x})]^2$ ). In general, obviously, the true density is not known, and one has to resort to recipes to choose the bandwidth that are derived based on some reasonable assumptions about the underlying distribution. A particularly simple heuristic for selecting the bandwidth is given by Silverman’s rule,<sup>46,47</sup>

$$\mathbf{H} = \left[ \frac{4}{N(D+2)} \right]^{\frac{2}{D+4}} \mathbf{\Sigma}, \quad (3)$$

where  $\mathbf{\Sigma}$  is the covariance of the entire dataset,  $N$  the number of data points and  $D$  the dimensionality.

### 2.3.1 Local Kernel Metric Analysis

Silverman’s rule (3) minimizes the MISE for a single multivariate normal distribution. For a general distribution, composed of many separate peaks, this generally results in an over-estimation of the bandwidth, and in loss of resolution unless an extremely large amount of data is available. As a possible solution, and to provide a mechanism to fine-tune the balance between resolution and statistical noise, we propose a simple strategy to localize the determination of the bandwidth and of the dimensionality of the data. Basically, the idea is to apply Silverman’s rule to subsets of the full dataset.

In order to estimate the optimal KDE bandwidth for a sample  $\mathbf{x}_i$ , we introduce weighting factors for each of the other sample points  $\mathbf{x}_j$  around it, that are computed from a spherical

Gaussian

$$u_{ij} = \exp \left[ -\frac{(\mathbf{x}_j - \mathbf{x}_i)^2}{2 \cdot \sigma_i^2} \right] N w_j / \sum_j w_j, \quad (4)$$

where  $\sigma_i$  is a localization factor whose choice we will discuss below. The weights-adjusted sample population for the selected point is computed as  $N_i = \sum_j u_{ij}$ . We find it convenient to introduce two possible approaches to determine the localization parameters  $\sigma_i$ . (i) in cases where one expects the spatial extent of clusters to be relatively homogeneous, one can choose a fixed localization window expressed as a fraction of the overall spatial extent of the dataset, assessed as the global covariance matrix of the data,  $\Sigma$ ; this can be achieved setting  $\sigma_i = f_s \sqrt{\text{Tr } \Sigma}$ . In this case, each localization region can contain a different weight-adjusted population  $N_i$ . (ii) In cases where one expects clusters with very different spreads, but similar populations, it might be more convenient to use a position-dependent localization window that is adjusted so that each region contains a prescribed fraction  $f_p$  of the total number of weighted data points. Each  $\sigma_i$  should then be adjusted iteratively until  $N_i \approx N f_p$ . A quantitative discussion of the relative merits of the two strategies is reported in the SI (Fig. S1).

The local weights from eq. (4) can be used, for instance, to estimate the local covariance  $\Sigma_i$ , around each data point. Each element of the covariance is estimated by

$$[\Sigma_i]_{kl} = \sum_{j=1}^N u_{ij} \cdot (x_{jk} - \bar{x}_k)(x_{jl} - \bar{x}_l) / N_i \quad (5)$$

where  $\bar{\mathbf{x}} = \sum_{j=1}^N u_{ij} \mathbf{x}_j / N_i$ . Computing the covariance of a subset of the points can exacerbate stability problems that are also present, in general, when the sampling is insufficient or when different degrees of freedom are strongly correlated. In these circumstances,  $\Sigma_i$  can be very ill-conditioned, and even have eigenvalues that are zero to within machine precision. This is a consequence of the fact that the usual estimator for the covariance  $\Sigma_i$  is a biased estimator of its inverse  $\Sigma_i^{-1}$ . This is a well-known problem that is typically addressed by introducing alternative estimators that are less strongly biased. Here we use the so-called



Oracle Approximating Shrinkage (OAS) estimator<sup>48</sup> that reads

$$\tilde{\Sigma}_i = (1 - \psi_i)\Sigma_i + \frac{\psi_i \text{Tr}(\Sigma_i)\mathbf{I}}{D} \quad (6)$$

where

$$\psi_i = \min \left[ 1, \frac{(1 - \frac{2}{D}) \text{Tr}(\Sigma_i^2) + \text{Tr}^2(\Sigma_i)}{(N_i + 1 - \frac{2}{D}) \text{Tr}(\Sigma_i^2) - \frac{\text{Tr}^2(\Sigma_i)}{D}} \right].$$

Furthermore, the eigenvalue spectrum of the local covariance matrix can be used to estimate an effective local dimensionality  $D_i$  based on the effective rank of  $\Sigma_i$ :<sup>49</sup>

$$D_i = \exp \left( - \sum_{k=1}^D \eta_k \log(\eta_k) \right) \quad (7)$$

where  $\eta_k = \lambda_k / \sum_{k=1}^D |\lambda_k|$  and  $\{\lambda_k\}$  is the eigenvalue spectrum of  $\Sigma_i$ .

Given the local covariance matrices and an estimate of the local dimensionality, one can introduce an expression for the optimal bandwidth matrices to be used in the KDE. Assuming that each local zone resembles a normal distribution, the optimal bandwidth for  $\mathbf{x}_i$  can finally be obtained as a localized version of Silverman's Rule (3):

$$\mathbf{H}_i = \left[ \frac{4}{N_i(D_i + 2)} \right]^{\frac{2}{D_i+4}} \tilde{\Sigma}_i. \quad (8)$$

Performing this analysis for each datum would entail poor scaling of the procedure with the total number of points. One can however exploit the definition of a sparse grid to accelerate greatly the computation, without changing the spirit of the localization strategy. First, one can compute the local bandwidth  $\mathbf{H}_i$  only for the grid points  $\mathbf{y}_i$ , and assign it to all the samples that belong to its Voronoi set, i.e.  $\mathbf{H}_j \equiv \mathbf{H}_i$ ,  $\tilde{\Sigma}_j \equiv \tilde{\Sigma}_i \quad \forall \mathbf{x}_j \in \mathcal{V}_i$ . Furthermore, the evaluation of eq. (5) can be accelerated by including the contribution from grid points beyond a reasonable cutoff distance using only the position of the grid  $\mathbf{y}_i$  and assigning to it a weight proportional to the total weight of points within its associated Voronoi polyhedron, rather

than summing over all sample points associated to it.

Finally, we note that in cases in which sampling is particularly irregular, it can happen that the bandwidth is smaller than the distance with the nearest neighbor of a grid point. Often these outliers result from insufficient grid size and/or non-Gaussian tails of the distribution, and should not generate additional clusters. To avoid this, we increase automatically the bandwidth to match the first-neighbor distance, but issue a warning to allow for manual inspection to determine whether the outlier is of some significance.

## 2.4 Identification of Motifs

After having estimated the density at the grid points  $P(\mathbf{y}_i)$ , one can proceed to use it to subdivide the distribution into several distinct clusters. As discussed above, we chose to identify clusters that represent recurring molecular patterns as maxima in the probability distribution, and to associate to each maximum all the grid points falling within its basin of attraction. In atomistic trajectories this construction has a profound physical interpretation: each identified maximum of the probability distribution can be associated with a free energy (meta-)stable minimum of the  $D$ -dimensional description of a group of atoms.

We perform a non-parametric clustering based on this idea using the Quick-Shift algorithm.<sup>50</sup> Other clustering algorithms could obviously be used - see the SI for a comparison of some of the most common methods.<sup>51-53</sup> Starting from a random grid point which has not been assigned yet to a cluster, one connects it to the nearest grid point that has a higher probability density, i.e.,  $\mathbf{y}_i$  is connected to  $\mathbf{y}_j$  such that

$$j = \underset{P(\mathbf{y}_i) > P(\mathbf{y}_j)}{\operatorname{argmin}} |\mathbf{y}_i - \mathbf{y}_j|. \quad (9)$$

The procedure can be interrupted based on a suitable stopping criterion, as discussed below. The final point is identified as a maximum, and tagged as the center  $\mathbf{z}_k$  of a cluster. All the points in the chain are tagged as belonging to the associated set  $\mathcal{Z}_k$ . One can then start

climbing from another unassigned point, and stop when the procedure encounters another maximum, or one of the points that have already been assigned to one of the  $\mathcal{Z}_k$  clusters, with which the current chain would then be merged. After all points have been traversed, the grid  $\mathcal{Y}$  will be partitioned into  $n$  disjoint sets, without the need of specifying *a priori* the number of clusters or their geometry. The stopping criterion for Quick-Shift is typically set by requiring that the Euclidean distance between the current point and the next one in the chain is below a set threshold  $\Delta$ . In line with the spirit of the local metric analysis we perform to obtain the bandwidth matrix for the KDE, we select the cutoff locally. At each stage in the procedure, the cutoff is chosen based on the covariance associated with the grid point that is being considered, i.e.

$$\Delta_i = \alpha \sqrt{\text{Tr } \tilde{\Sigma}_i}. \quad (10)$$

This choice adapts the cutoff to the local spread of the data, and is consistent with the Gaussian assumption that underlies the localization procedure. For a multivariate Gaussian, it would cluster together two points that are drawn at random from the distribution. If needed, the values of  $\Delta_i$  can be further adjusted by a multiplicative factor  $\alpha$ , to fine-tune the resolution of the clustering procedure. In the examples we used to benchmark this method we found only small changes in the final clustering upon scaling the cutoff by  $\pm 10\%$ .

#### 2.4.1 Gaussian Mixture Models

After having determined the number and position of modes of the distribution, we fit a Gaussian Mixture Model (GMM) to the data. This step is useful as it provides a simple interpretation of the density in terms of a number of separate modes, making it possible to develop fingerprints for the different molecular motifs that have a transparent probabilistic

interpretation. The probability distribution is thus fitted to a sum of  $n$  multivariate Gaussians

$$\hat{P}(\mathbf{x}) = \sum_{k=1}^n p_k G(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad (11)$$

Each Gaussian, associated with a weight  $p_k$ , is defined as

$$G(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D \det \boldsymbol{\Sigma}}} e^{-(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})/2} \quad (12)$$

where  $\boldsymbol{\Sigma}$  is its covariance matrix and  $\boldsymbol{\mu}$  its mean position. Rather than fitting the Gaussian parameters with an expectation-maximization algorithm, we exploit the fact that we know the number  $n$  and modes  $\mathbf{z}_k$  of clusters. We set the mean of the Gaussian cluster to the mode of the cluster, and estimate the covariance with the usual expression:

$$\begin{aligned} \boldsymbol{\mu}_k &= \mathbf{z}_k, \quad p_k = \sum_{\mathbf{y} \in \mathcal{Z}_k} P(\mathbf{y}) / \sum_{\mathbf{y} \in \mathcal{Y}} P(\mathbf{y}), \\ \bar{\mathbf{y}}_k &= \sum_{\mathbf{y} \in \mathcal{Z}_k} \mathbf{y} P(\mathbf{y}) / \sum_{\mathbf{y} \in \mathcal{Z}_k} P(\mathbf{y}) \\ \boldsymbol{\Sigma}_k &= \sum_{\mathbf{y} \in \mathcal{Z}_k} (\mathbf{y} - \bar{\mathbf{y}}_k)(\mathbf{y} - \bar{\mathbf{y}}_k)^T P(\mathbf{y}) / \sum_{\mathbf{y} \in \mathcal{Z}_k} P(\mathbf{y}). \end{aligned} \quad (13)$$

The reason for introducing this additional step, after having already obtained a non-parametric clustering by quick-shift, is that a GMM lends itself quite naturally to a probabilistic interpretation. Given a configuration associated with the fingerprints  $\mathbf{x}$ , the expression

$$\hat{P}_k(\mathbf{x}) = p_k G(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) / \left( \zeta + \hat{P}(\mathbf{x}) \right) \quad (14)$$

corresponds to the probability that such configuration belongs to the  $k$ th cluster. Eqn. (14) can therefore be used to introduce some smooth, fuzzy Probabilistic Motif Identifiers (PMI) that constitute a data-driven definition of a molecular pattern – such as the hydrogen bond<sup>28,29</sup> or the accumulation of charge around an excess proton.<sup>30</sup> The “background” parameter  $\zeta$  –

that defaults to zero and should in any case be set to a very small value – serves to provide a more physical description of outlier configurations. In practice, configurations for which all Gaussian densities are below  $\zeta$  are considered to be new, unclassified states that had not been sampled properly in the initial dataset.

### 2.4.2 Periodic data

When working with descriptors that are periodic in nature (e.g. angles and dihedrals) the probabilistic description should be adapted to account for the non-Euclidean geometry of the space. Von Mises distributions<sup>54,55</sup> are the equivalent of a Gaussian on a circle, and can be used to describe periodic data, as they are smooth across the boundaries. The multivariate extension of a von Mises distribution, however, cannot be normalized analytically, which makes it hard to use it for the KDE step in our procedure. Furthermore, when the bandwidth is negligible with respect to the periodicity, a Gaussian kernel computed while using a minimal image convention in defining distances between points is virtually indistinguishable from a von Mises distribution. For this reason, we use multivariate Gaussian kernels in the KDE step, also along periodic directions.

When determining the GMM that underlies our fingerprints, however, one cannot assume that the covariance associated with each cluster is small with respect to the periodicity of the pattern space. Given the difficulties with normalizing a multivariate von Mises distribution, we use a product of one-dimensional distributions to construct basis functions for the GMM that represent each cluster, i.e. we use

$$G(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\kappa}) = \prod_i^D \frac{e^{\kappa_i \cos(x_i - \mu_i)}}{2\pi I_0(\kappa)}, \quad (15)$$

in lieu of Eq. (12), where  $I_0$  is the modified Bessel function of order zero. The mean value and weight for each cluster are determined according to (13), whereas  $\kappa_i$  are obtained using the conventional estimators for the concentration parameter of a one-dimensional von Mises

distribution.<sup>56</sup>

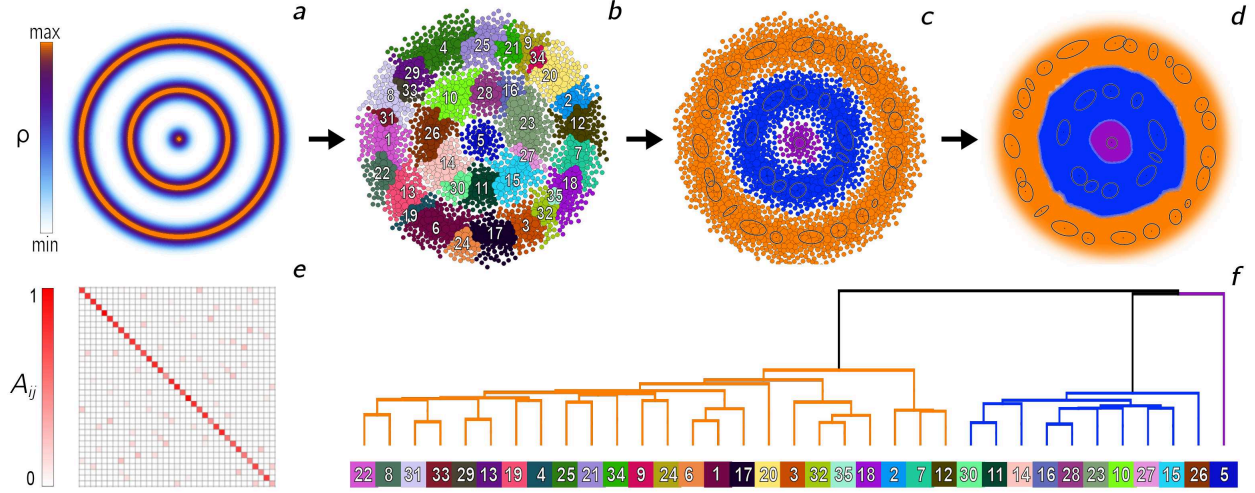


Figure 2: (a) Radially-symmetric probability density function that corresponds to concentric circles. Clusters before (b) and after (c) merging and the PMIs corresponding to the final macroclusters (d). Panel (f) represents the dendrogram resulting from the agglomerative clustering based on the cluster stability matrix  $\mathbf{A}$  (e) and using a single-linkage strategy for the merging.

## 2.5 Error Assessment

Computing the *absolute* MISE requires knowledge of the underlying probability distribution. It is however possible to estimate the statistical error associated with a given estimate, which can be useful to fine-tune the KDE parameters and to decide on the statistical significance of clusters that are identified in later stages of the PAMM analysis. Bootstrapping provides a very general and well-established approach to infer the statistical error in a distribution, when its analytic form is not known.<sup>57</sup> Bootstrapping relies on the analogy existing between the population and the sample drawn from it and consists in re-sampling with replacement a large number of sets from the given data, in order to build empirically an estimate of the probability distribution related to a certain statistical estimate. In this specific case, we exploit bootstrapping to generate  $N_{BS}$  independent samples of the KDE,  $P^{(m)}(\mathbf{y}_i)$ . From these, one can estimate the standard error  $\delta P(\mathbf{y}_i)$  associated with the KDE at each grid point.

The bootstrapping procedure is not only useful to get an estimate of the statistical error in the KDE. By performing the (deterministic) clustering procedure we discussed above on the  $m$ -th bootstrapped estimate of  $P(\mathbf{y}_i)$ , one can obtain clusters  $\mathcal{Z}_k^{(m)}$  that reflect the statistical fluctuations of the KDE. The comparison between the bootstrapped clusters and those obtained on the straightforward KDE can then be used to compute indicators of how “stable” the clustering procedure can be considered. To do so, one can first compute

$$\begin{aligned} Q &= \sum_i P(\mathbf{y}_i), & Q_k &= \sum_{\mathbf{y} \in \mathcal{Z}_k} P(\mathbf{y}), \\ Q_k^{(m)} &= \sum_{\mathbf{y} \in \mathcal{Z}_k^{(m)}} P(\mathbf{y}), & Q_{j|k}^{(m)} &= \sum_{\mathbf{y} \in \mathcal{Z}_k \cap \mathcal{Z}_j^{(m)}} \frac{P(\mathbf{y})}{Q_k^{(m)}}, \end{aligned} \quad (16)$$

and then introduce

$$A_{ij} = \frac{1}{N_{\text{BS}} \sqrt{Q_i Q_j}} \sum_m \sum_k Q_k^{(m)} Q_{i|k}^{(m)} Q_{j|k}^{(m)}. \quad (17)$$

For each bootstrapping run, this expression determines what is the probability that – taking one of the bootstrap cluster at random, and drawing two sample points from them – one would be part of the reference cluster  $i$ , and one of the reference cluster  $j$ , renormalized over the probabilities of clusters  $i$  and  $j$ . The diagonal elements  $A_{ii}$  report on how robust is the determination of the  $i$ -th cluster. If the  $i$ -th cluster appears identical in each bootstrapping run,  $A_{ii}$  takes a value of one, which becomes smaller if in one or more of the iterations the cluster is split in multiple clusters. Off-diagonal terms report on how “fuzzy” are the borders of the clusters. If no bootstrapping run generates a cluster that overlaps with both  $\mathcal{Z}_i$  and  $\mathcal{Z}_j$ ,  $A_{ij}$  would take a value of zero, which increases if the clusters get merged in some of the runs, or if some of the  $\mathcal{Z}_k^{(m)}$  include points from both clusters.

### 2.5.1 Cluster Association and Non-Gaussian Patterns.

The cluster stability matrix  $A_{ij}$  from eq. (17), can also be used to perform an additional “meta-clustering” step, that suggests ways to group together some of the clusters identified in

the previous steps of PAMM, based on the notion that they were separated due to statistical error rather than because they correspond to separated free-energy basins. We attempted two approaches, that provide satisfactory and similar results, but differ in the underlying interpretation. One possibility is to choose a threshold value for the adjacency matrix, and find the connected components of the associated graph. This approach corresponds to a sort of “flooding” scheme, in which clusters that are above a prescribed level of fuzziness are merged at once.

Alternatively, one can proceed with a hierarchical clustering procedure,<sup>58</sup> which can also be represented in a tree-like plot which helps interpreting the relations between different clusters.<sup>59</sup> Based on the adjacency matrix one can define a distance between clusters as  $d_{ij} = -\log(A_{ij}/\sqrt{A_{ii}A_{jj}})$ . The pair of clusters which are closest is merged first, after which the merging is repeated iteratively until a single cluster remains. The cluster hierarchy can be represented as a binary tree, in which the vertical position of the branching point correspond to the distance between the leaves. Different strategies exist – and can be tried to improve the resolving power of the method – to define a distance between merged clusters. In this work we always use Ward’s minimum-variance prescription,<sup>60</sup> unless otherwise specified. This second strategy is more consistent with an interpretation in which fuzzy clusters correspond to clustering errors, and performs as little merging as possible to achieve the desired number of clusters, or degree of separation.

Besides making the clustering procedure more robust, this merging step is also useful to address the presence of strongly non-Gaussian features in fingerprint space. Even though the KDE and Quick-Shift algorithms are fully non-parametric, at many steps in our protocol we invoked the assumption that data can be (locally) described by multivariate Gaussian distributions. As a specific, and rather extreme, example of non-Gaussian behavior, let us consider the distribution depicted in Fig. 2, corresponding to three concentric rings. Figure 2 demonstrates how, in the presence of non-Gaussian clusters, the partitioning of the data by PAMM is highly unstable, leading to an adjacency matrix showing considerable overlap



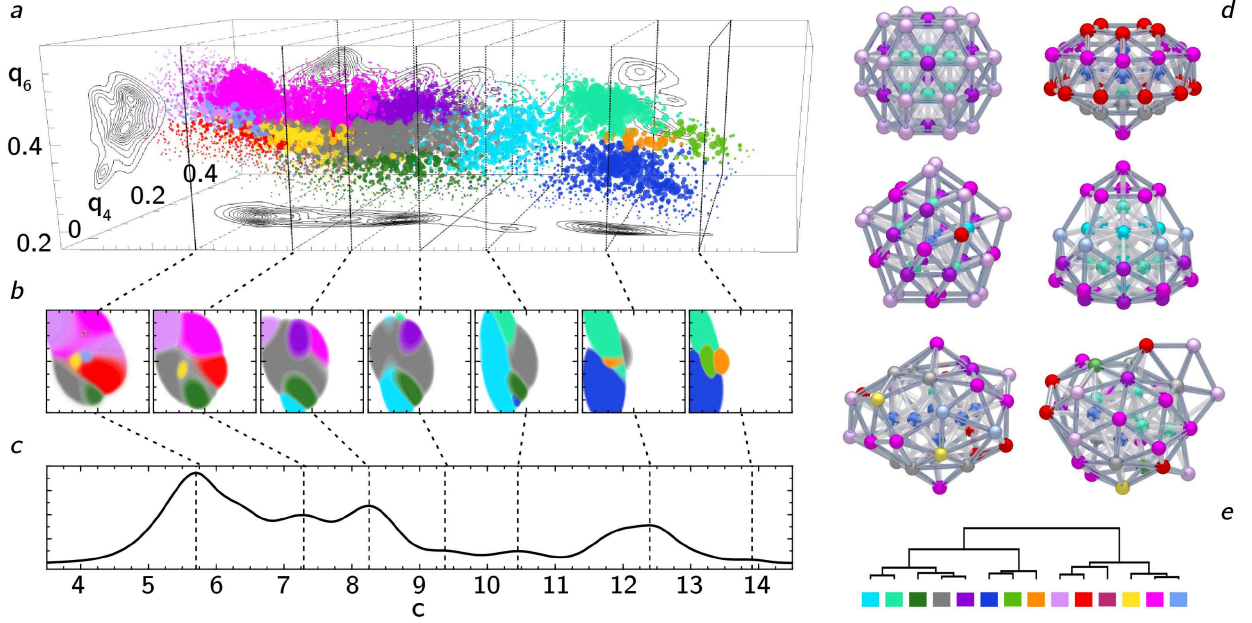


Figure 3: A PAMM analysis of local environments in a  $\text{LJ}_{38}$  cluster simulated at  $T^* = 0.18$ . (a) The probability distribution in fingerprint space, that include a smooth coordination number  $c$ , and the Steinhardt order parameters  $q_4$  and  $q_6$ . Grid points have an area proportional to the KDE of the probability density, and are colored according to the quick-shift clustering. Marginal probabilities are also drawn as contour plots. (b) Slices of the probabilistic motif indicators that result from the GMM built on the PAMM clusters. Each cluster is indicated with the corresponding solid color when its PMI is equal to 1, while the opacity linearly decreases to 0 when the value of the PMI is 0. (c) The probability distribution for  $c$  only. (d) Representative configurations of the  $\text{LJ}_{38}$  clusters, with atoms colored according to the dominant PMI. (e) Binary tree representation of the hierarchical clustering based on the adjacency matrix of the PAMM clusters.

between different clusters. Hierarchical clustering, illustrated using a dendrogram in Fig. 2e, shows clearly that there are three “macro-cluster” that correspond to the rings, while Gaussian features within each ring are clearly detected as being strongly connected. It is worth noting that, once different Gaussian clusters have been joined based on the adjacency matrix, it is easy to develop non-Gaussian fingerprints, by simply summing over all GMM fingerprints associated with each macro-cluster  $\mathcal{M}$

$$\hat{P}_{\mathcal{M}}(\mathbf{x}) = \sum_{k \in \mathcal{M}} p_k G(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) / \left( \zeta + \hat{P}(\mathbf{x}) \right). \quad (18)$$

A demonstration of this non-Gaussian fingerprint is also depicted in Fig. 2d.

### 3 Probabilistic Motif Identifiers

The main use case for PAMM involves the data-driven identification of local environments that characterize the state of atoms or molecular fragments, e.g. the secondary structure in a biomolecule or the crystal structure of a solid. For this reason, we will first demonstrate our approach by determining such probabilistic motif identifiers (PMIs) in two very different cases: a Lennard-Jones cluster, that is an archetypical example of the problems encountered in nanomaterials, followed by the example of a 16 residue long hairpin peptide (GB1), which we discuss as a simple but realistic example of the applications to biological systems.

#### 3.1 Local Motifs of a Lennard-Jones Cluster

The LJ<sub>38</sub> cluster has been used very often as a benchmark of minimization algorithms, free-energy techniques, and structure-recognition methods, because its potential energy landscape contains a very deep, narrow enthalpic minimum corresponding to a truncated *fcc* lattice, and a broad basin containing a multitude of defective structures corresponding to icosahedral symmetry.<sup>42,61–63</sup> Here we used data from the  $T = 0.18T^*$  replica of a long parallel tempering trajectory,<sup>42</sup> that contains structures that are representative of both the solid phases and liquid-like, highly defective configurations.

Several of the possible environments (e.g. corner, edge, facet and core atoms) can be roughly identified by the number of nearest neighbors, that can be characterized based on purely radial information. Here we use the coordination number  $c(i)$  for particle  $i$ , that we define as

$$c(i) = \sum_{j \neq i}^N \frac{1}{\exp\left(\frac{r_{ij} - r_c}{\gamma}\right) + 1} \quad (19)$$

where  $\mathbf{r}_{ij}$  indicates the vector between particle  $i$  and  $j$  and  $r_{ij} = |\mathbf{r}_{ij}|$  is the Euclidean distance between those two atoms. In order for the Fermi function to count the number of first neighbors, we set  $r_c = 1.45\sigma$  and  $\gamma = 0.2\sigma$ .

While neighbor counts can be very effective in identifying motifs in the minimum-energy

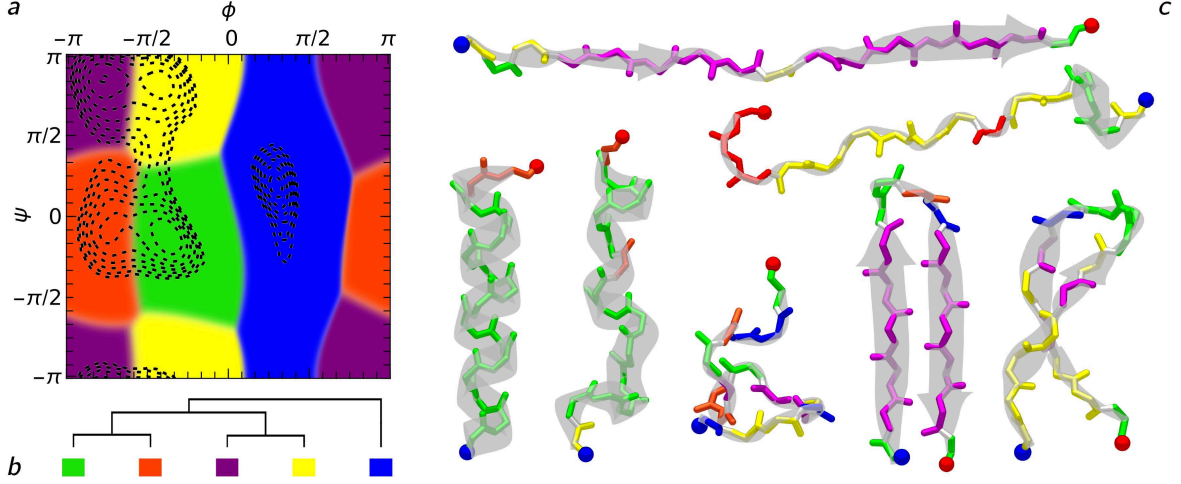


Figure 4: (a) PMIs for the backbone dihedrals of a  $\beta$ -hairpin are shown as a function of the Ramachandran angles. Different regions are colored based on the cluster that is associated with the dominant PMI. The underlying probability distribution is shown as black contours, equally spaced on a logarithmic scale. (b) Dendrogram representation of the hierarchical clustering of the adjacency matrix of the PAMM clusters. (c) A few representative structures of the molecules are shown where the residues are colored according to the dominant PMI. The color code corresponds to that used in panel (a).

structures, finite-temperature simulations can be considerably fuzzier. Furthermore, the coordination number cannot distinguish between bulk environments from a *fcc* lattice and the core of an icosahedral motif. To address these problems, and to demonstrate the behavior of PAMM in a non-trivial example, we supplemented  $c(i)$  with angular information. We used the local bond order parameters (or Steinhardt order parameters)<sup>64</sup>  $q_4(i)$  and  $q_6(i)$ , which are known to be able to distinguish body center cubic (*bcc*), face centered cubic (*fcc*) or hexagonal close packed (*hcp*) crystal lattices.<sup>65–67</sup>

We use the definition

$$q_l(i) = \sqrt{\frac{4\pi}{2l+1} \sum_{m=-l}^l |q_{lm}(i)|^2} \quad (20)$$

in which the complex vector  $q_{lm}(i)$  is defined as

$$q_{lm}(i) = \frac{1}{c(i)} \sum_{j \neq i}^N \frac{Y_{lm}(\mathbf{r}_{ij})}{\exp\left(\frac{r_{ij}-r_c}{\gamma}\right) + 1} \quad (21)$$

The functions  $Y_{lm}(\mathbf{r}_{ij})$  are the spherical harmonics and the loop runs over all particles, since the Fermi function singles out contributions from just the first coordination shell. Since  $c(i)$ ,  $q_4(i)$  and  $q_6(i)$  have very different ranges, when combining the different order parameters to give a 3D descriptor of the environments, we centered and scaled them so that each component has unit variance.

Figure 3 demonstrates the outcome of a PAMM analysis for the finite-temperature trajectory. A point based approach with a smoothing parameter  $f_{\text{points}} = 0.02$  was used to compute the KDE, while for the clustering step the scaling factor  $\alpha$  was set to 1. PAMM can recognize motifs based on both coordination number and angular correlations based on the three-dimensional joint probability density (Fig. 3a) – identifying several more clusters than it would be possible based on  $c(i)$  alone (Fig. 3c). The PMIs for the different motifs (Fig. 3b) can be used to recognize the motifs in each snapshot of the simulation (Fig. 3d), and could be used, e.g., to bias a molecular dynamics simulation which triggers transitions between different structures. In this relatively simple case clusters are clear-cut, robust to sub-sampling and to variations of the PAMM parameters, and with an approximately Gaussian shape. Even though an agglomerative meta-clustering step is therefore not necessary, we did compute the cluster stability matrix, and generated the associated hierarchical clustering tree (Fig. 3e). Inspection of the binary tree is insightful, showing that individual clusters can be grouped in two main branches, corresponding to surface and bulk atoms. As we will also see in other cases, it appears that the hierarchical merging of the PAMM clusters can be interpreted much in the same way as a disconnectivity graph,<sup>68,69</sup> reporting on the relations between different basins in pattern space.

### 3.2 Local Motifs of a $\beta$ -hairpine Peptide

Secondary structure in proteins is a textbook example of how molecular motifs can behave as building blocks of complex supramolecular structures. Many algorithms exist to identify secondary structure patterns (SSPs), that are based on the identification of hydrogen

bonds.<sup>70,71</sup> More recently, it has been shown that secondary structure conformations can be classified solely on the basis of backbone dihedrals of a protein.<sup>72,73</sup> As a demonstration of our classification approach to realistic periodic data, we have applied our method to analyze the data of the backbone dihedral pairs  $(\phi, \psi)$  from a replica exchange simulation of a 16-residue C-terminal fragment of the immunoglobulin binding domain B1 of the Streptococcus protein G in explicit solvent (GB1, amino acids sequence Ace-GEWTYDDATKTFTVTE-Nme) (for further details of the simulation see Ardevol et al.<sup>74</sup>). For each residue and each frame of the trajectory we computed the backbone dihedral angles  $\phi$  and  $\psi$ , and performed a PAMM analysis. The underlying KDE was estimated by using a point-based KDE smoothing  $f_p = 0.15$ , and step-scaling  $\alpha = 1.0$  for the subsequent Quick-Step clustering. The resulting PMIs result in a partitioning of the Ramachandran  $\phi - \psi$  plot<sup>75</sup> in 5 regions, that correspond roughly to  $\beta$  sheets,  $\alpha$  helices, turns, etc. (see Fig. 4), and that are clearly associated with local probability maxima in the KDE (shown as black contours in Fig. 4a).

Even though the clusters are identified as von Mises modes, with a single basis function assigned to each of them, it is clear that the PMI correspond very accurately to the partitioning of the probability density in basins of attraction, with the transition zones between two PMIs following closely the dividing surface between basins. In Fig. 4b we also show a few reference structures selected from the trajectory. The aminoacids in the backbone have been colored based on the dominant PMI to which they are associated.

It is easy to recognize the PMIs associated with well known secondary structure elements by comparing reference structures to the PAMM partitioning of the Ramachandran plot. One can clearly identify, e.g.  $\alpha$ -helices or the antiparallel  $\beta$ -sheets that are abundant in the simulation data for the GB1 fragment. Chains of dihedrals that correspond to a  $\beta$ -sheet conformation are also seen in extended structures. We also find several instance of turn-type T1 motifs,<sup>73</sup> that in our case concentrate in the unstructured portions of the polypeptide.

Meanwhile, it is clear that there is not a 1:1 correspondence between PMIs and “traditional” secondary structure motifs. Some of the PMIs correspond to portions of the Ramachandran

plot that are traditionally assigned to polyproline II (PPII) helices and left-handed helices, yet we could not identify a significant presence of stable structures associated with those motifs. Rather, extended structures and distorted  $\beta$  strands are associated with the PPII region, while left-handed helical patterns appear, together with many other PMIs, within disordered, “random-coil” configurations. This observation highlights the potential of a data-driven approach to the definition of molecular patterns. Well-separated clusters in fingerprint space are recognized even though they do not appear clearly by just inspecting the trajectory in search for well-established structural motifs. This agnostic behavior could be very useful, for instance, in rationalizing the behavior of intrinsically-disordered proteins,<sup>76,77</sup> or to study polypeptides in unusual environments, such as at inorganic interfaces or in combination with synthetic polymers. Furthermore, an automated probability analysis allows one to extend the definition of pattern space by combining e.g. several backbone dihedrals, or by combining dihedrals and H-bonding indicators. This procedure could give rise to more precise identification of secondary-structure patterns, and will be the subject of future research. Finally, the hierarchical clustering of the five PAMM motifs (Fig. 4b) shows another example of how the adjacency matrix built by a bootstrapping analysis reflects the structure of the free-energy landscape of patterns, with the linkage distance corresponding roughly to the free-energy barrier between basins.

## 4 Global Structure Classification

PAMM clustering of low-dimensional fingerprint spaces allowed us to demonstrate some of the improved features of this implementation, including periodic descriptors, adjacency-matrix determination and hierarchical “meta-clustering”. We now move on to demonstrate the applicability of our method to problems with higher dimensionality. Rather than using higher-dimensional descriptors of local motifs, we consider the extreme case of attempting a global classification of the meta-stable states of the cluster and the polypeptide that we

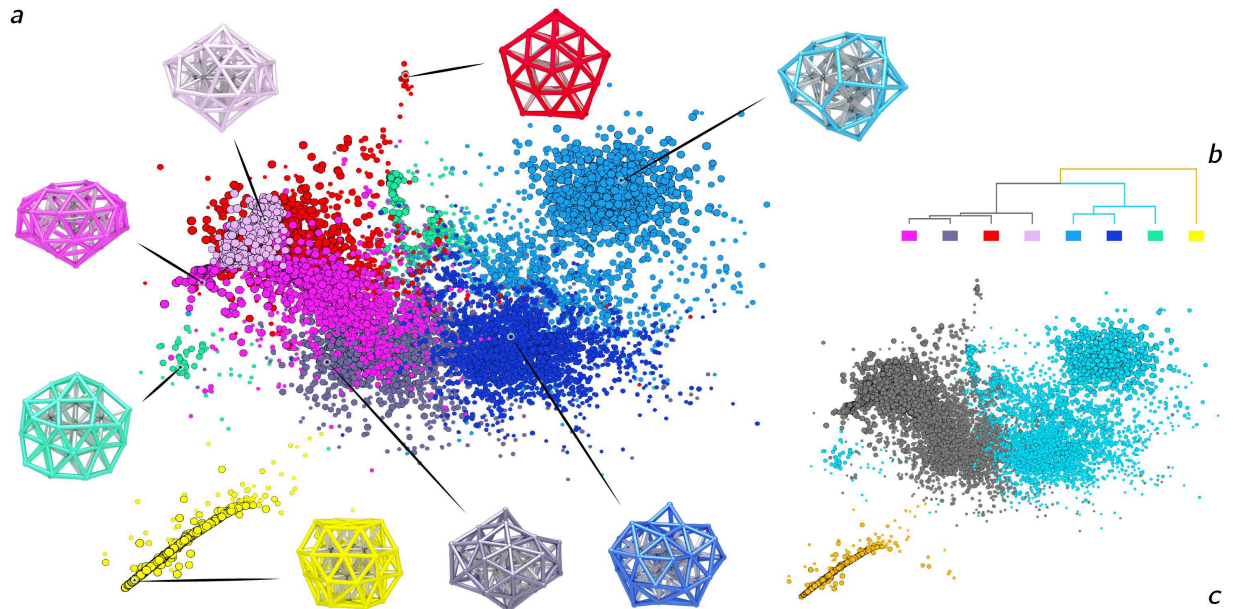


Figure 5: PAMM classification of configurations extracted from a simulation of  $\text{LJ}_{38}$  at  $T^* = 0.18$ . (a) The clusters generated from the first stage of PAMM using the parameters  $f_p = 0.05$  for the (point-based) KDE smoothing and  $\alpha = 1$  for the scaling of the Quick-Shift cutoff. Configurations from the trajectory are represented using a two-dimensional sketch-map representation. The size of points reflects the probability density. Points are colored according to the cluster to which they belong. (b) The initial clusters are merged according to the hierarchical clustering procedure. (c) Sketch-map representation colored according to the macro-clusters, together with a snapshot representative of the highest-probability region for each motif.

discussed in the previous Section. These problems are intrinsically high-dimensional, and do not simply involve an overdetermined description of a relatively low-dimensional problem. In order to simplify the interpretation of our results, we combined PAMM clustering with a two-dimensional representation of the configurational landscape. Among the many non-linear dimensionality reduction algorithms available,<sup>78–82</sup> we used sketch-map,<sup>83,84</sup> a method that was already used successfully to study these systems.<sup>41,74</sup> We show that the combination of the two methods helps greatly overcoming some of their limitations, and understanding the relation between different structures in complex systems.

## 4.1 Structural classification for $\text{LJ}_{38}$

$\text{LJ}_{38}$  is one of the classical examples of simple clusters exhibiting a major structural transition, between a truncated-octahedral structure and a defective icosahedron. Since we consider a temperature close to the melting point, also liquid-like configurations should be present in the trajectory.

In order to identify different structures, we use the same 15-dimensional descriptors based on a smooth histogram of coordination numbers that were introduced in Ref. 41, and use them both as the basis for a PAMM analysis and as the input for the construction of a sketch-map representation, for which we used the same parameters and reference map as in Ref. 41.

We decided not to tune the fingerprints that had already been used in the literature so as to test the stability of PAMM-based clustering when dealing with sub-optimal inputs. For instance, some of the histogram bins have near-zero variance, which would yield singular bandwidth matrices if we did not stabilize the estimator with the OAS. Furthermore, the sharp cutoff function used in defining the coordination number histogram leads to the presence of a large number of basins, corresponding to minute differences in the structures. However, by using a point-based localization with  $f_p = 0.05$ , we implicitly consider a probabilistic model in which each cluster contains at least 5% of the data. As a result, the smearing parameter and the Quick-Shift threshold are large enough that many of these minute clusters coalesce, leaving just 8 motifs detected (Figure 5a).

The number and type of clusters, however, depend rather sensitively on the PAMM parameters. If one inspects the structural features that are associated with the initial PAMM clusters (see Figure 5a) it becomes clear that PAMM clustering does identify portions of configuration space that are clearly distinct, corresponding e.g. to different defective “gemstone” structures, to the truncated octahedron, and to liquid-like structures with different kinds of surface geometry. A dendrogram representation of the hierarchical clustering based on the adjacency matrix (Figure 5b) shows that some of the clusters are very sensitive



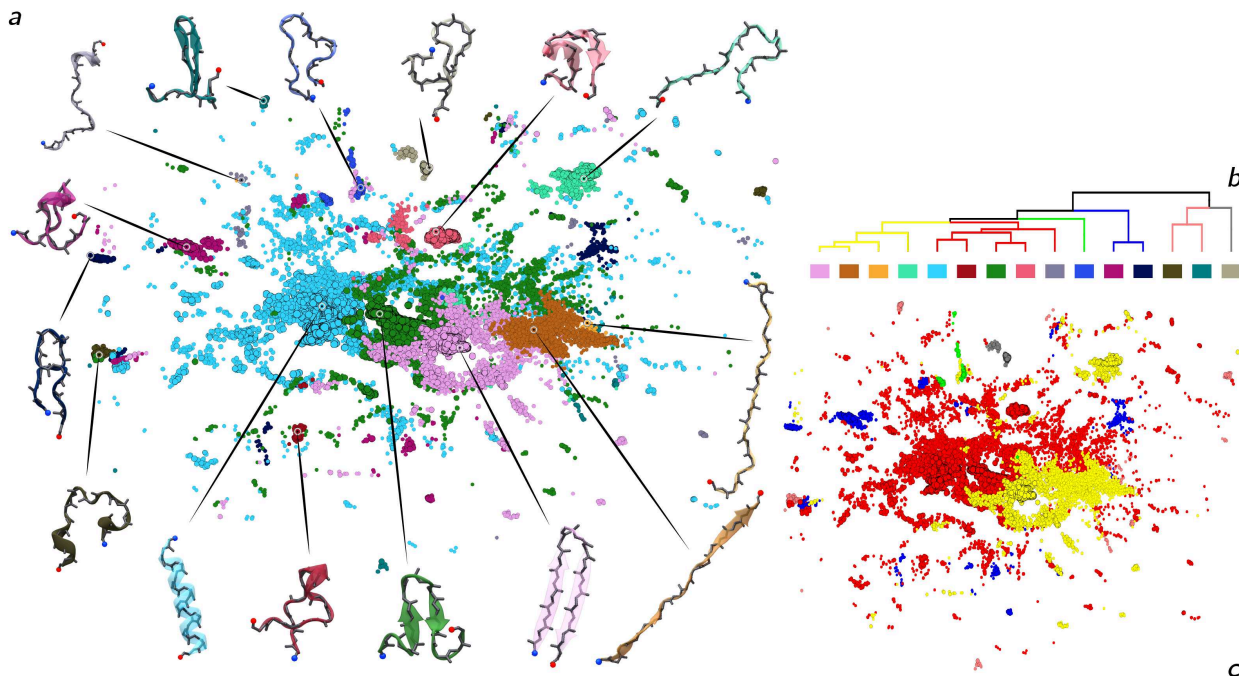


Figure 6: PAMM classification of configurations extracted from a simulation of the GB1 hairpin fragment. (a) Clusters generated in the first stage of PAMM using the parameters 0.1 for the (point-based) KDE smoothing and  $\alpha = 1$  for the scaling of the Quick-Shift cutoff. Configurations from the trajectory are represented using a two-dimensional sketch-map representation. The size of points reflects the probability density. Points are colored according to the cluster to which they belong. (b) The initial clusters are merged according to the hierarchical clustering procedure. (c) Sketch-map representation colored according to the macro-clusters, together with a snapshot representative of the highest-probability region for each conformation.

to statistical noise, and resulting in high adjacency. One can identify three very stable meta-clusters, that are represented in (Figure 5c). PAMM recognizes that the most prominent features of the free-energy landscape comprise liquid-like states, more or less disordered icosahedral fragments and – even more clearly separated – the truncated octahedron. Once again it is remarkable how the adjacency matrix, despite being constructed just as a heuristic indicator of cluster stability, reflects the connectivity of the free-energy landscape. Although this far we were not able to reveal a formal connection, this aspect will be the subject of further investigation.

One final consideration concerns the relationship between the clustering and the dimensionality reduction approaches to describe complex atomistic systems. A sketch-map

representation does offer a more intuitive, bird’s eye view of the landscape. However, in order to achieve such a high level of coarse graining, the dimensionality reduction algorithm introduces considerable distortion, including also discontinuities in the projection.<sup>41,84,85</sup> For this reason, it is important to perform the clustering step in the high-dimensional space to avoid having a classification which is biased by sketch-map artifacts. For instance, Figure 5 shows clearly that one of the clusters identified by PAMM is split in two by the projection, without a continuous path connecting the fragments on the map. Thus, combining clustering and dimensionality-reduction approaches might provide a strategy to compensate for the shortcomings of the two methods, and obtain deeper understanding of the structural features of the system.

## 4.2 Structural classification for a $\beta$ -hairpine Peptide

The structural landscape for the GB1 oligopeptide provides another suitable benchmark for the application of PAMM to the clustering of high-dimensional data. This  $\beta$ -hairpin fragment has been studied extensively by metadynamics,<sup>86</sup> and has been used to demonstrate the comparison between wild type and mutant proteins using sketch-map.<sup>74</sup> Such analysis revealed a rugged free-energy landscape, containing many metastable states including a helical configuration, several mis-folded hairpin configurations as well as the native fold.

As in the case of the the LJ<sub>38</sub> cluster, we used a simple-minded choice of high-dimensional descriptors – namely, the 30 backbone dihedrals – as the input representation. This choice minimizes the bias on the clustering procedure, but implies that configurations that differ by minor details (e.g. the configuration of the terminal aminoacids) are considered as distinct structures. Indeed, a global classification of the GB1 peptide produces a multitude of clusters (Figure 6a), and varies considerably depending on the clustering details and bootstrapping resampling. Hierarchical cluster merging, however, simplifies considerably this picture (Figure 6b), making it possible to group states that are associated with a helical configuration, the native fold, and a few misfolded hairpin configurations (Figure 6c). Clustering also

highlights the limitations of the sketch-map projection, that gives a contiguous representation of the high-probability clusters (helix and native hairpin) but scatters higher-free-energy states at the periphery of the map.

## 5 Conclusions

We presented a density-based clustering approach to analyze the outcome of atomistic simulations in search for the essential atomic-scale patterns that characterize the behavior of materials and molecules. The algorithm generalizes the recently-introduced PAMM method to periodic, high-dimensional and/or sparsely sampled data, combining several state-of-the-art techniques to guarantee robust, reliable and efficient clustering. Crucially, the method we introduce performs repeated clustering attempts on top of re-sampled distributions constructed by bootstrapping, and uses these to construct an adjacency matrix that characterizes the stability and the overlaps between clusters. Combined with a hierarchical “meta-clustering” and a binary tree representation this further analysis serves two purposes. First, it provides a way to improve the quality of PAMM clusters – by merging non clear-cut motifs and allowing the representation of strongly non-Gaussian modes in the probability. Second, it makes it possible to recognize the relations between the fine-grained details of the free-energy landscape and the main basins on a coarser scale. In all the examples we considered, the hierarchical clustering reflects qualitatively the structure of the free energy, resembling a disconnectivity graph. It might be possible to formulate the construction of the adjacency matrix in a way that makes this analogy quantitative, an endeavor which will be the subject of future research.

We demonstrate that this improved PAMM algorithm identifies coordination environments in a LJ<sub>38</sub> cluster and secondary-structure motifs in a 16-residues  $\beta$ -hairpin peptide. Although our method is geared towards recognizing *local* molecular patterns – and more principled approaches exist that are designed to identify *global* conformational states<sup>87–90</sup> – we also show

it can cluster overall structures in these two systems. The combination of PAMM clustering, binary-tree merging of the motifs and sketch-map representation of high-dimensional free energy landscapes provide multiple insights into the behavior of complex atomistic systems, overcoming some of the limitations of the methods, such as the presence of discontinuities in sketch-map projections. Development code to perform a PAMM analysis is available from an on-line repository <sup>1</sup>, and the data we used for demonstration purposes is provided in the SI.

The probabilistic motif identifiers associated with each cluster can be used as fuzzy, smoothly varying collective variables in biased molecular dynamics schemes to accelerate sampling and reconstruct the underlying free-energy landscape. The Gaussian Mixture Model associated with the PMI could also be used as an ansatz for the target probability distribution in a variational-sampling scheme,<sup>91</sup> with the populations of the clusters as the variational parameters. Data-driven approaches to recognize the building blocks of complex materials constitute a necessary ingredient to assist the analysis and interpretation of large-scale atomistic trajectories, and provide a natural representation of free-energy landscapes in terms of the most stable configurations that can be used to accelerate configurational sampling.

## Acknowledgement

The Authors would like to thank Gareth Tribello for insightful discussion and help implementing PMIs in PLUMED. PG acknowledges funding from the MPG-EPFL center for molecular nanoscience, and from the NCCR MARVEL, funded by the Swiss National Science Foundation.. MC and RHM were supported by the European Research Council under the European Union’s Horizon 2020 research and innovation programme (grant agreement no. 677013-HBMAP).

---

<sup>1</sup><http://cosmo-epfl.github.io>

## References

- (1) Fischer, C. C.; Tibbetts, K. J.; Morgan, D.; Ceder, G. Predicting crystal structure by merging data mining with quantum mechanics. *Nature materials* **2006**, *5*, 641.
- (2) Marzari, N. Materials modelling: The frontiers and the challenges. *Nature materials* **2016**, *15*, 381–382.
- (3) Lindorff-Larsen, K.; Maragakis, P.; Piana, S.; Shaw, D. E. Picosecond to Millisecond Structural Dynamics in Human Ubiquitin. *The Journal of Physical Chemistry B* **2016**, *120*, 8313–8320.
- (4) Zhao, G.; Perilla, J. R.; Yufenyuy, E. L.; Meng, X.; Chen, B.; Ning, J.; Ahn, J.; Gronenborn, A. M.; Schulten, K.; Aiken, C.; Zhang, P. Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature* **2013**, *497*, 643–646.
- (5) Agrawal, A.; Choudhary, A. Perspective: Materials informatics and big data: Realization of the fourth paradigm of science in materials science. *APL Materials* **2016**, *4*, 053208.
- (6) Bereau, T.; Andrienko, D.; Kremer, K. Research Update: Computational materials discovery in soft matter. *APL Materials* **2016**, *4*, 053101.
- (7) Curtarolo, S.; Hart, G. L.; Nardelli, M. B.; Mingo, N.; Sanvito, S.; Levy, O. The high-throughput highway to computational materials design. *Nature materials* **2013**, *12*, 191.
- (8) Phillips, C. L.; Littlewood, P. Preface: Special Topic on Materials Genome. *APL Materials* **2016**, *4*, 053001.
- (9) Pizzi, G.; Cepellotti, A.; Sabatini, R.; Marzari, N.; Kozinsky, B. AiiDA: automated interactive infrastructure and database for computational science. *Computational Materials Science* **2016**, *111*, 218–230.

- (10) Chodera, J. D.; Singhal, N.; Pande, V. S.; Dill, K. A.; Swope, W. C. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *The Journal of chemical physics* **2007**, *126*, 04B616.
- (11) de Groot, B. L.; Daura, X.; Mark, A. E.; Grubmüller, H. Essential dynamics of reversible peptide folding: memory-free conformational dynamics governed by internal hydrogen bonds. *Journal of molecular biology* **2001**, *309*, 299–313.
- (12) Andrec, M.; Felts, A. K.; Gallicchio, E.; Levy, R. M. Protein folding pathways from replica exchange simulations and a kinetic network model. *Proceedings of the National Academy of Sciences of the United States of America* **2005**, *102*, 6801–6806.
- (13) Singhal, N.; Snow, C. D.; Pande, V. S. Using path sampling to build better Markovian state models: predicting the folding rate and mechanism of a tryptophan zipper beta hairpin. *The Journal of chemical physics* **2004**, *121*, 415–425.
- (14) Huan, T. D.; Mannodi-Kanakkithodi, A.; Ramprasad, R. Accelerated materials property predictions and design using motif-based fingerprints. *Physical Review B* **2015**, *92*, 014106.
- (15) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical review letters* **2012**, *108*, 058301.
- (16) Faber, F. A.; Lindmaa, A.; von Lilienfeld, O. A.; Armiento, R. Machine Learning Energies of 2 Million Elpasolite (A B C 2 D 6) Crystals. *Physical Review Letters* **2016**, *117*, 135502.
- (17) Behler, J. Perspective: Machine learning potentials for atomistic simulations. *The Journal of Chemical Physics* **2016**, *145*, 170901.

- (18) De, S.; Musil, F.; Ingram, T.; Baldauf, C.; Ceriotti, M. Mapping and classifying molecules from a high-throughput structural database. *Journal of Cheminformatics* **2017**, *9*, 6.
- (19) Sparks, T. D.; Gaultois, M. W.; Oliynyk, A.; Brgoch, J.; Meredig, B. Data mining our way to the next generation of thermoelectrics. *Scripta Materialia* **2016**, *111*, 10–15.
- (20) Rost, B.; Sander, C. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins: Structure, Function, and Bioinformatics* **1994**, *19*, 55–72.
- (21) Ballester, P. J.; Mitchell, J. B. A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics* **2010**, *26*, 1169–1175.
- (22) Cheng, J.; Baldi, P. A machine learning information retrieval approach to protein fold recognition. *Bioinformatics* **2006**, *22*, 1456–1463.
- (23) Dale, J. M.; Popescu, L.; Karp, P. D. Machine learning methods for metabolic pathway prediction. *BMC bioinformatics* **2010**, *11*, 15.
- (24) Haranczyk, M.; Sethian, J. A. Automatic structure analysis in high-throughput characterization of porous materials. *Journal of chemical theory and computation* **2010**, *6*, 3472–3480.
- (25) Carr, D. A.; Lach-hab, M.; Yang, S.; Vaisman, I. I.; Blaisten-Barojas, E. Machine learning approach for structure-based zeolite classification. *Microporous and Mesoporous Materials* **2009**, *117*, 339–349.
- (26) Schütt, K.; Glawe, H.; Brockherde, F.; Sanna, A.; Müller, K.; Gross, E. How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. *Physical Review B* **2014**, *89*, 205118.

- (27) Kayala, M. A.; Baldi, P. Reactionpredictor: Prediction of complex chemical reactions at the mechanistic level using machine learning. *Journal of chemical information and modeling* **2012**, *52*, 2526–2540.
- (28) Gasparotto, P.; Ceriotti, M. Recognizing molecular patterns by machine learning: An agnostic structural definition of the hydrogen bond. *J. Chem. Phys.* **2014**, *141*, 174110.
- (29) Gasparotto, P.; Hassanali, A. A.; Ceriotti, M. Probing Defects and Correlations in the Hydrogen-Bond Network of ab Initio Water. *J. Chem. Theory Comput.* **2016**, *12*, 1953–1964.
- (30) Rossi, M.; Ceriotti, M.; Manolopoulos, D. E. Nuclear Quantum Effects in H<sup>+</sup> and OH–Diffusion along Confined Water Wires. *The Journal of Physical Chemistry Letters* **2016**, *7*, 3001–3007.
- (31) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.
- (32) Bartók, A. P.; Gillan, M. J.; Manby, F. R.; Csányi, G. Machine-learning approach for one- and two-body corrections to density functional theory: Applications to molecular and condensed water. *Phys. Rev. B* **2013**, *88*, 054104.
- (33) Sadeghi, A.; Ghasemi, S. A.; Schaefer, B.; Mohr, S.; Lill, M. A.; Goedecker, S. Metrics for measuring distances in configuration spaces. *J. Chem. Phys.* **2013**, *139*, 184118.
- (34) Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **2011**, *134*.
- (35) Pietrucci, F.; Andreoni, W. Graph Theory Meets Ab Initio Molecular Dynamics: Atomic Structures and Transformations at the Nanoscale. *Phys. Rev. Lett.* **2011**, *107*, 085504.
- (36) Kitao, A.; Go, N. Investigating protein dynamics in collective coordinate space. *Current Opinion in Structural Biology* **1999**, *9*, 164 – 169.



- (37) Branduardi, D.; Gervasio, F. L.; Parrinello, M. From A to B in free energy space. *The Journal of Chemical Physics* **2007**, *126*, 054103.
- (38) Nüske, F.; Keller, B. G.; Pérez-Hernández, G.; Mey, A. S.; Noe, F. Variational approach to molecular kinetics. *Journal of chemical theory and computation* **2014**, *10*, 1739–1752.
- (39) Tiwary, P.; Berne, B. Spectral gap optimization of order parameters for sampling complex molecular systems. *Proceedings of the National Academy of Sciences* **2016**, *113*, 2839–2844.
- (40) Rosenkrantz, D. J.; Stearns, R. E.; Philip M. Lewis, I. An Analysis of Several Heuristics for the Traveling Salesman Problem. *SIAM Journal on Computing* **1977**, *6*, 563–581.
- (41) Ceriotti, M.; Tribello, G. A.; Parrinello, M. Demonstrating the Transferability and the Descriptive Power of Sketch-Map. *J. Chem. Theory Comput.* **2013**, *9*, 1521–1532.
- (42) Ceriotti, M.; Bussi, G.; Parrinello, M. Colored-Noise Thermostats à la Carte. *J. Chem. Theory Comput.* **2010**, *6*, 1170–1180.
- (43) Prabhakaran, S.; Raman, S.; Vogt, J. E.; Roth, V. *Lecture Notes in Computer Science*; Springer Berlin Heidelberg, 2012; pp 458–467.
- (44) Ester, M.; Kriegel, H.; Sander, J.; Xiaowei, X. *A density-based algorithm for discovering clusters in large spatial databases with noise*; 1996.
- (45) Rodriguez, A.; Laio, A. Machine learning. Clustering by fast search and find of density peaks. *Science (New York, N.Y.)* **2014**, *344*, 1492–1496.
- (46) Scott, D. *Multivariate density estimation : theory, practice, and visualization*; Wiley: New York, 1992.
- (47) Bowman, A. W.; Azzalini, A. *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations (Oxford Statistical Science Series)*; Oxford University Press, 1997.

- (48) Chen, Y.; Wiesel, A.; Eldar, Y. C.; Hero, A. O. Shrinkage Algorithms for MMSE Covariance Estimation. *IEEE Transactions on Signal Processing* **2010**, *58*, 5016–5029.
- (49) Roy, O.; Vetterli, M. The effective rank: A measure of effective dimensionality. 15th European Signal Processing Conference. 2007; pp 606–610.
- (50) Vedaldi, A.; Soatto, S. In *Computer Vision – ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12–18, 2008, Proceedings, Part IV*; Forsyth, D., Torr, P., Zisserman, A., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2008; pp 705–718.
- (51) Carreira-Perpinan, M. A. Mode-finding for mixtures of Gaussian distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2000**, *22*, 1318–1323.
- (52) others,, et al. Some methods for classification and analysis of multivariate observations. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. 1967; pp 281–297.
- (53) others,, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd. 1996; pp 226–231.
- (54) Mardia, K. V.; Jupp, P. E. Distributions on spheres. *Directional Statistics* **2000**, 159–192.
- (55) Mardia, K. V.; Hughes, G.; Taylor, C. C.; Singh, H. A Multivariate Von Mises Distribution with Applications to Bioinformatics. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* **2008**, *36*, 99–109.
- (56) Sra, S. A short note on parameter approximation for von Mises-Fisher distributions: and a fast implementation of  $I_s(x)$ . *Computational Statistics* **2012**, *27*, 177–190.
- (57) Efron, B. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics* **1979**, *7*, 1–26.

- (58) Murtagh, F.; Contreras, P. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2012**, *2*, 86–97.
- (59) De, S.; Musil, F.; Ingram, T.; Baldauf, C.; Ceriotti, M. Mapping and classifying molecules from a high-throughput structural database. *Journal of Cheminformatics* **2017**, *9*, 6.
- (60) Ward Jr, J. H. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* **1963**, *58*, 236–244.
- (61) Wales, D. J.; Doye, J. P. K. Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms. *The Journal of Physical Chemistry A* **1997**, *101*, 5111–5116.
- (62) Wales, D. J.; Miller, M. A.; Walsh, T. R. Archetypal energy landscapes. *Nature* **1998**, *394*, 758–760, Copyright - Copyright Macmillan Journals Ltd. Aug 20, 1998; Last updated - 2012-11-14; CODEN - NATUAS.
- (63) Prtay, L. B.; Bartk, A. P.; Csny, G. Efficient Sampling of Atomic Configurational Spaces. *The Journal of Physical Chemistry B* **2010**, *114*, 10502–10512, PMID: 20701382.
- (64) Steinhardt, P. J.; Nelson, D. R.; Ronchetti, M. Bond-orientational order in liquids and glasses. *Phys. Rev. B* **1983**, *28*, 784–805.
- (65) Moroni, D.; ten Wolde, P. R.; Bolhuis, P. G. Interplay between Structure and Size in a Critical Crystal Nucleus. *Phys. Rev. Lett.* **2005**, *94*, 235703.
- (66) Coasne, B.; Jain, S. K.; Naamar, L.; Gubbins, K. E. Freezing of argon in ordered and disordered porous carbon. *Phys. Rev. B* **2007**, *76*, 085416.
- (67) Ogata, S. Monte Carlo simulation study of crystallization in rapidly supercooled one-component plasmas. *Phys. Rev. A* **1992**, *45*, 1122–1134.

- (68) Becker, O. M.; Karplus, M. The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. *J. Chem. Phys.* **1997**, *106*, 1495.
- (69) Mortenson, P. N.; Evans, D. A.; Wales, D. J. Energy landscapes of model polyalanines. *J. Chem. Phys.* **2002**, *117*, 1363.
- (70) Frishman, D.; Argos, P. Knowledge-based protein secondary structure assignment. *Proteins: Structure, Function, and Bioinformatics* **1995**, *23*, 566–579.
- (71) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577–2637.
- (72) Hollingsworth, S. A.; Lewis, M. C.; Berkholz, D. S.; Wong, W.-K.; Karplus, P. A.  $(\phi, \psi)_2$  Motifs: A Purely Conformation-Based Fine-Grained Enumeration of Protein Parts at the Two-Residue Level. *Journal of Molecular Biology* **2012**, *416*, 78 – 93.
- (73) Nagy, G.; Oostenbrink, C. Dihedral-Based Segment Identification and Classification of Biopolymers I: Proteins. *Journal of Chemical Information and Modeling* **2014**, *54*, 266–277.
- (74) Ardevol, A.; Tribello, G. A.; Ceriotti, M.; Parrinello, M. Probing the Unfolded Configurations of a  $\beta$ -Hairpin Using Sketch-Map. *J. Chem. Theory Comput.* **2015**, *11*, 1086–1093.
- (75) Ramachandran, G. N.; Ramakrishnan, C.; Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *Journal of molecular biology* **1963**, *7*, 95–99.
- (76) Tompa, P.; Davey, N. E.; Gibson, T. J.; Babu, M. M. A million peptide motifs for the molecular biologist. *Molecular cell* **2014**, *55*, 161–169.
- (77) Receveur-Br  chot, V.; Durand, D. How random are intrinsically disordered proteins?

- A small angle scattering perspective. *Current Protein and Peptide Science* **2012**, *13*, 55–75.
- (78) Kruskal, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **1964**, *29*, 1–27.
- (79) Jolliffe, I. T. *Principal component analysis*; Springer, 1986; pp 115–128.
- (80) Abdi, H.; Williams, L. J. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics* **2010**, *2*, 433–459.
- (81) Coifman, R. R.; Lafon, S.; Lee, A. B.; Maggioni, M.; Nadler, B.; Warner, F.; Zucker, S. W. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences of the United States of America* **2005**, *102*, 7426–7431.
- (82) Maaten, L. v. d.; Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research* **2008**, *9*, 2579–2605.
- (83) Ceriotti, M.; Tribello, G. A.; Parrinello, M. From the Cover: Simplifying the representation of complex free-energy landscapes using sketch-map. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 13023–8.
- (84) Tribello, G. A.; Ceriotti, M.; Parrinello, M. Using sketch-map coordinates to analyze and bias molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 5196–201.
- (85) Comitani, F.; Rossi, K.; Ceriotti, M.; Sanz, M. E.; Molteni, C. Mapping the conformational free energy of aspartic acid in the gas phase and in aqueous solution. *J. Chem. Phys.* **2017**, *146*, 145102.
- (86) Bonomi, M.; Branduardi, D.; Gervasio, F. L.; Parrinello, M. The Unfolded Ensemble and Folding Mechanism of the C-Terminal GB1  $\beta$ -Hairpin. *J. Chem. Am. Soc.* **2008**, *130*, 13938–13944.

- (87) Weber, M.; Rungsarityotin, W.; Schliep, A. *Perron cluster analysis and its connection to graph partitioning for noisy data*; Konrad-Zuse-Zentrum für Informationstechnik Berlin, 2004.
- (88) Deuffhard, P.; Weber, M. Robust Perron cluster analysis in conformation dynamics. *Linear algebra and its applications* **2005**, *398*, 161–184.
- (89) Röblitz, S.; Weber, M. Fuzzy spectral clustering by PCCA+: application to Markov state models and data classification. *Advances in Data Analysis and Classification* **2013**, *7*, 147–179.
- (90) Chodera, J. D.; Noé, F. Markov state models of biomolecular conformational dynamics. *Current opinion in structural biology* **2014**, *25*, 135–144.
- (91) Valsson, O.; Parrinello, M. Variational Approach to Enhanced Sampling and Free Energy Calculations. *Phys. Rev. Lett.* **2014**, *113*, 090601.